# Discussion

*Thomas R. Belin*
*University of California -- Los Angeles, School of Public Health*

The papers presented in this session (Hostetter 1994; Winkler, 1994; Steel and Konschnik 1994; Kovacevic and Liu, 1994; Griffin, 1994) represent a range of applications and theoretical development in the arenas of record linkage and statistical matching. By "record linkage" or "exact matching," we are referring to problems where the goal is to combine information from separate databases on the same individuals; the papers by Hostetter, Winkler, and Steel and Konschnik fall into this category. By "statistical matching," we are referring to problems where the goal is to combine information from separate databases on people who look similar to one another; the paper by Kovacevic and Liu and the paper by Griffin fall into this category.

The questions that have guided much of record-linkage research might be characterized as follows:

(1) How can we optimize exact matching procedures?

  (1a) What is the best configuration of factors that are at the discretion of the operator of the computer matching program for producing good matching results?

  (1b) What factors in computer matching offer the most promise for obtaining further gains in matching accuracy?

(2) How should exact-matched data sets be analyzed to reflect accurately any uncertainty about matching?

Experience suggests that gains in accuracy are frequently available by taking advantage of additional information in the matching procedure, but beyond a certain point gains in accuracy are elusive, in part because factors interact. Belin (1993) explores both (1a) and (1b), invoking principles of statistical design and analysis. A framework for evaluating record linkage procedures follows from viewing the computer matcher as a "black box," with data and factors at the program operator's control being thought of as input and the accuracy of designated matches being the output. Examples of results that emerged from an analysis of census data were (i) adding a variable like marital status to information such as name, address, birth date, age, race, and sex did not always improve the accuracy of matching, and (ii) methods for accommodating imperfect agreement in name or address frequently led to substantial improvements in matching accuracy, but not always.

Regarding question (2), Scheuren and Winkler (1993) explicitly consider the uncertainty in subsequent statistical analyses that should follow from uncertainty about matching. There is very little attention to this idea elsewhere in the literature. To elaborate on a possible direction for future work, suppose as in Steel and Konschnik's paper, we are interested in the total payroll for employers in the United States. Steel and Konschnik reported that $34.2 billion shifted from the non-employer file to the employer file after their computer matching. But suppose for a group of records, there is some uncertainty about whether they belong to one file or the other. An idea for dealing with this scenario would be to create plausible realizations of the data file (perhaps 2 or 3 or 5) based on estimated probabilities of belonging to one file or the other. From these realizations, one could calculate the mean payroll shift in each, as well as the variability in payroll shift across realizations, which could then be used in developing an inference about the quantity of interest.

Similarly, we can ask in the context of statistical matching:

(1) How can we optimize statistical matching procedures?

(1a) What algorithms should be used?

(1b) What factors should be controlled?

(2) How should statistically-matched data sets be analyzed to reflect accurately any uncertainty about matching?

The papers by Griffin and by Kovacevic and Liu are primarily focussed on (1a). A reference in the literature that goes straight to the heart of question (2) is Rubin (1986).

To elaborate on these ideas with some notation, suppose, as in the context of Kovacevic and Liu that we want to know the association between tax changes and changes in consumer retail spending. We can envision the following setup:

| Database | Variables |
|---|---|
| 1 | Matching variables (X), tax data (Y) |
| 2 | Matching variables (X), spending data (Z). |

Statistical matching requires either an assumption about the conditional association of Y and Z given X, conditional independence being a special case bound to be violated frequently in practice, or external information about the joint distribution of Y and Z given X, e.g., from census data, a smaller study, or another source of earlier data that may still appear to be relevant.

Below, I summarize a number of items that could be labeled "impressions." There might be some difference of opinion from some quarters on one or another of these points, but they still may merit consideration as further work in statistical matching is pursued.

## ■ Impressions:

(1) Statistically matched data sets that do not reflect a range of plausible assumptions about conditional associations may not be useful for policymaking. (I use the term "may not" instead of "will not" because in some settings a point estimate from statistical matching might serve as an "objective" or agreed-upon approach to deciding an issue on which there are irreconcilable nontechnical differences. But in general, having a point estimate without a sense of its uncertaintywould not inform the policymaking process very much.)

(2) In realistic applications, there will be many conditional associations about which there is no information from available data.

(3) In general, the more auxiliary information on conditional associations that is used in statistical matching, the better.

(4) There will always be an assumption in statistical matching that pairs of variables are conditionally independent, given some (possibly large) set of variables.

(5a) If there are census data available on the association between two variables Y and Z, then there is no need to perform statistical matching; on the other hand,

(5b) If one is working with a smaller study or an earlier study as an external source of information, it does not seem as wise to impose a constraint on a conditional association as it does to reflect some uncertainty about the conditional association.

(6) A singly imputed statistically matched data set will generally be inadequate to reflect uncertainty in secondary analyses (Rubin 1986).

(7) I am less enthusiastic about optimization methods, such as simulated annealing for implementing statistical matching, than about iterative simulation methods, such as data augmentation (Tanner and Wong, 1987) or Gibbs sampling (Gelfand and Smith,1990) that could be used to generate imputations.

Both record linkage and statistical matching address questions of considerable interest whose answers we do not know. A final impression is that both would be more widely used if progress can be made on the very difficult problem of accurately reflecting uncertainty in secondary analyses.

## ■ References


Belin, T.R. (1993), "Evaluation of Sources of Variation in Record Linkage through a Factorial Experiment," *Survey Methodology*, 19, 13-29.

Gelfand, A.E., and Smith, A.F.M. (1990), "Sampling-Based Approaches to Calculating Marginal Densities," *Journal of the American Statistical Association*, 85, 398-409.

Griffin, R.A. (1994), "An Application of Stochastic Optimization to Combine Two Files," *Proceedings of the Section on Survey Research Methods, American Statistical Association*, to appear.

Hostetter, S.C. (1994), "Linking Individuals on a Capital Gains Panel for Tax Policy Analysis," *Proceedings of the Section on Survey Research Methods, American Statistical Association*, to appear.

Kovacevic, M.S., and Liu, T.P. (1994), "Statistical Matching of Survey Data Files: An Empirical Study," *Proceedings of the Section on Survey Research Methods, American Statistical Association*, to appear.

Rubin, D.B. (1986), "Statistical Matching Using File Concatenation with Adjusted Weights and Multiple Imputations," *Journal of Business and Economic Statistics*, 4, 87-94.

Scheuren, F., and Winkler, W.E. (1993), "Regression Analysis of Data Files that are Computer Matched," *Survey Methodology*, 19, 39-58.

Steel, P.M., and Konschnik, C.A. (1994), "Administrative Record Matching for the 1992 Economic Censuses," *Proceedings of the Section on Survey Research Methods, American Statistical Association*, to appear.

Tanner, M.A., and Wong, W.H. (1987), "The Calculation of Posterior Distributions by Data Augmentation" (with discussion), *Journal of the American Statistical Association*, 82, 528-550.

Winkler, W.E. (1994), "Advanced Methods for Record Linkage," *Proceedings of the Section on Survey Research Methods, American Statistical Association*, to appear. ■